Supplementary Material for

Beyond Mono to Binaural: Generating Binaural Audio from Mono Audio with Depth and Cross Modal Attention

Kranti Kumar Parida[1] Siddharth Srivastava[2] Gaurav Sharma[1,3]

[1] IIT Kanpur [2] CDAC Noida [3] TensorTour Inc.

{kranti, grv}@cse.iitk.ac.in, siddharthsrivastava@cdac.in

1. Ablation

We provide further ablations for the outputs of the decoders. As mentioned in Fig. 2 of the main manuscript, our approach has three decoders each one predicting magnitude, STFT and phase of the difference of both channels individually. We use the STFT prediction branch for obtaining the final output binaural audio. We also experiment with a different architecture where there is a single decoder instead of three for predicting the output. We provide the results for different architectural variations in Tab. 1.

In Tab. 1, we give two different variants of the decoder architecture. In the first block of the table, we report the result for the case where there are three individual decoders. In the second block, instead of using three decoders we use a single decoder for obtaining the output. We note here that for both the cases we can add all the four different losses (Sec. 3.5 in main manuscript). But for the second case as we are not predicting magnitude and phase separately, we obtain them from the predicted STFT and use the same for magnitude, phase and reconstruction loss calculation. Hence, the output for second case is only the STFT prediction. We report the results for *split-1* of modified FAIR-Play dataset after 50 epochs.

We make two conclusions here. i) The addition of separate magnitude and phase prediction improves the performance (row-1 vs. row-4 in Tab. 1). We observe that there is significant improvement in performance of STFT, ENV from $1.301, 0.163$ to $1.171, 0.156$ with single decoder and three individual decoders respectively. We hypothesize that the primary reason for this are the very different mathematical functions for calculation of magnitude and phase as described in Sec 4.1 of main manuscript. ii) Adding individual Magnitude and Phase subnetworks regularizes the training process. As we are predicting magnitude and phase individually in the first case, we can also obtain the output by combining the predicted magnitude and phase from the respective networks. We give the results for both the cases in Tab. 1 (row1 vs. row2). When we obtain the output from STFT branch, we get a performance of 1.171, 0.156 for STFT and ENV respectively, whereas combining predictions from Magnitude and Phase gives a performance of 1.267 and 0.162 respectively. This shows that using the predictions from STFT branch give better performance. However, using explicit magnitude and phase loss is beneficial as its removal degrades the performance to 1.206 and 0.158 (row3 in Tab. 1). This observation confirms that Magnitude and Phase subnetworks act as regularizers even though the prediction from these branches do not add directly over the STFT branch.

Approach	$\mathcal{L}$	$\mathcal{L}_{mag}$	$\mathcal{L}_{phs}$	$\mathcal{L}_{rec}$	STFT (↓)	ENV (↓)
STFT	✓	✓	✓	✓	1.171	0.156
Mag-Phs	✓	✓	✓	✓	1.267	0.162
STFT only	✓	✗	✗	✗	1.206	0.158
STFT (w/ mag-phs)	✓	✓	✓	✓	1.301	0.163

Table 1. **Comparison of different output setting.** The first block of results (rows 1,2) are for three individual decoders and the second block (rows 3,4) are for those using a single decoder only. We observe that adding all the losses with individual decoder and taking output directly from STFT gives the best performance. ↓ indicates lower value is better.

2. Qualitative Videos

We give four qualitative videos obtained using our approach. We request the readers to look at the results available at our project page `https://krantiparida.github.io/projects/bmonobinaural.html`. We request the readers to use a high quality headphone and use both left and right speakers, to be able to appreciate the binaural audio. The first two files contain single sound producing source. In first video, the sound source is present towards the left and hence the audio is dominant in the left audio channel. In the second example, the sound is present towards the right side of the scene. Hence the audio is

dominant in the right audio channel. We can also observe in this example that as the source move towards the centre, the predicted audio also follows the trajectory. This shows that our approach is able to model even the subtle variations in the input. In third example, there are two sources in the scene, one on the left and other on the right. Our network successfully produces the binaural audio where the audio for each of the source goes predominantly to the corresponding channels. We show a limitation of our approach in the final example, where the two sound sources produce similar sounds resulting in a relatively inferior binaural audio. However, the generated binaural audio is still perceptually better than the mono audio. Hence our approach can further be improved if we integrate some form of source separation knowledge/prior into the model, which we consider as a promising future direction.